Title

Disk Failure Prediction in Heterogeneous Environments

Authors

Carlos, A
Rincón, C
Pâris, Jehan-François
et al.

Publication Date

2017-07-01

DOI

10.23919/spects.2017.8046776

Peer reviewed

Disk Failure Prediction in Heterogeneous Environments

Carlos A. Rincón C.*†, Jehan-François Pâris*, Ricardo Vilalta*, Albert M. K. Cheng* and Darrell D. E. Long‡
*Department of Computer Science, University of Houston, TX, Houston, USA.
†Department of Computer Science, Universidad del Zulia, Maracaibo, Venezuela.
‡Department of Computer Science, University of California, Santa Cruz, CA, USA.

***Abstract*—Recent studies have shown the benefits o f using SMART attributes to predict disk failures in homogeneous populations of disks from the same make and model. We address here the case of data centers with more heterogeneous disk populations, such as the ones described in the BackBlaze datasets, and propose to build global disk failure predictors that would apply to disks of all makes and models. Our first c hallenge was the large number of SMART parameters that were missing for most makes and models in many disk instances of our dataset. As a result, we had to discard the SMART attributes that were missing in at least 90 percent of the disks, which left us with 21 SMART attributes. We then applied a Reverse Arrangement Test to these attributes to select the strongest disk failure indicators.**

We investigated three different machine learning models (Decision Trees, Neural Networks, and Logistic Regression) using the 2015 BackBlaze data to train and validate our predictors. Our best model was a decision tree that identified t rue f ailure events among the disks that tested positive for at least one of our failure indicators. We then used the 2016 BackBlaze data to evaluate its performance. Our results show that our decision tree identifies at least 52 percent of all disk failures and makes nearly all its predictions several days ahead: no more than 2.45 percent of the predicted failures occur within one day or two of the prediction.

Finally, we compared the performance of our predictor with those of the RAIDShield and the original BackBlaze predictor. We found out that RAIDShield could predict at most 18 percent of disk failures, that is, 34 percent fewer failures than our decision tree while the BackBlaze predictor predicted 60 percent of disk failures but generated 4 to 5 false alarms per correct prediction.

Keywords: hard disk failure prediction, SMART attributes, machine learning, data-driven simulation

I. Introduction

Magnetic disk drives offer the most cost-effective way to store large amounts of data. At the same time, they happen to be the least reliable component of modern computers, mostly because they include fast moving parts [1], [2]. As a result, all disk-based storage systems must include provisions for preventing data losses, such as replication and omission correcting codes [3], [4].

A more proactive approach can supplement these provisions. Rather than waiting for disks to fail, we could try to predict which ones are the most likely to fail by periodically sampling their SMART attributes, several of which are positively correlated with an increased risk of disk failure. This would let us save ahead of time the contents of the suspected disks, thus reducing the risk of data loss. Several recent studies show the viability of the approach [5], [6], [7]; their sole limitation resides in their focus on homogeneous populations of disks from the same make and model. In reality, many data centers include disks from multiple makes and models as new disks are constantly added to replace failed ones or to expand their storage capacity. One well-known instance is the BackBlaze disk farm [8]. To address this issue, we decided to focus our study on an exemplar of a heterogeneous disk population, namely, that described in BackBlaze quarterly hard drive reliability reports [8]. We used the 2015 data to train our predictors and data from 2016 to evaluate their performance, treating all disks equally regardless of their specific makes and models. Given the high number of SMART attributes that were missing for many disks, we used a two-step approach. First, we identified six SMART attributes that were the most correlated with an impending disk failure for our whole disk population. Second, we use machine learning to build a decision tree discrimination between true failure predictions and false alarms among the small minority of disks that tested positive for at least one of our failure indicators.

The results presented here illustrate the limits and the benefits of the approach. We found out that we could predict around 52 percent of all disk failures. The main culprit is the lack of standardization among the various SMART attributes that are reported by disks of different makes, which greatly reduced the number of attributes that were good overall failure indicators for our disk population.

We also compared the performance of our predictor with those of the RAIDShield [9] and the original BackBlaze predictor. We found out that RAIDShield could predict at most 18 percent of disk failures, that is, 34 percent fewer failures than our decision tree. In addition, RAIDShield could only achieve this result by setting its detection threshold at the lowest possible level, producing many more false alarms than correct predictions. The BackBlaze predictor would predict 60 percent of disk failures, that is, 8 percent more than our predictor but generated four to five false alarms per correct prediction.

II. Related Works

Xu *et al.* [5] used a health status assessment parameter (instead of a binary parameter) based on the SMART values to introduce a novel method based on recurrent neural networks to predict disk failures. In 2014, researchers from the same group (Li [7]) used classification and regression trees to predict disk failures based on the same health status assessment parameter. Both works applied the same methodology based on the reverse

SummerSim-SPECTS, 2017 July 9-12, Bellevue, Washington, USA;

arrangement test, rank-sum test, and z-scores to select the SMART parameters. They also used the same three datasets, each consisting of disks from the same manufacturer, to train, validate, and score their models. The results from both papers showed at least 95 percent accuracy for the best test case. However, due to the homogeneity of their datasets and the missing information about the number of models per dataset, their solutions will not work with high accuracy for data centers with disks from other manufacturers.

Botezatu *et al.* [6] presented a method based on decision trees to predict the necessity of a disk replacement based on the SMART parameter values. This approach utilized statistical techniques to automatically select the SMART parameters that correlate with a disk replacement. The results showed a 98 percent accuracy using as dataset 30,000 disks from two manufacturers (monitored over a period of 17 months). The proposed solution will not be applicable to a real-world data center because the researchers only used a limited dataset (one model per manufacturer for the training and validation process, and one model per manufacturer for the scoring process) to train, validate, and score the proposed model.

Zhu *et al.* [10] applied support vector machines and neural networks to predict disk failures. They used a dataset from a real data center consisting of 23,395 identical model drives. The results showed that the highest failure detection rate (95 percent) was achieved with the neural network model while the lowest false alarm ratio (0.03 percent) was achieved with the support vector machine model. Even though the authors used a real-world dataset, the distribution of the disks in the dataset (only one model) makes these results not applicable to heterogeneous environments.

Other studies, among them [11], [12], [13], and [14], applied different machine learning and statistical solutions to predict disk failures based on the SMART parameter values. In all four cases, their authors only considered population consisting of the same make and often the same models.

More recently, Jing Li *et al.* [15] proposed to measure the effectiveness of disk failure prediction models using the migration rate and the mismigration rate that result from the model predictions. Their work shows that warning time is as important as the accuracy of the model when measuring the performance of disk failure predictor.

We propose here to develop a different methodology to train, validate and score a disk failure prediction model by using a real-world dataset that includes disks from different manufacturers and models. This should help us understand the limitations of the selected machine learning techniques to predict disk failures in such environments.

III. Disk Failure Prediction

A. The Datasets

To work with a real-world dataset from a heterogeneous environment, we decided to use data from BackBlaze [8]. They measure 88 SMART parameters (44 raw and 44 normalized) per disk per day. To train and validate our model, we use the data from 2015 (17,509,251 records), while for scoring our model we use the data from 2016 (24,472,345 records).

The 2015 dataset from BackBlaze includes 62,898 drives. The fraction of drives that failed in this period is 2.27 percent (1,428 drives). Table I shows the distribution in terms of the manufacturer, the number of drives, and the number of models for the 2015 dataset.

TABLE I: BackBlaze 2015 Dataset Distribution

Manufacturer	Number of Disks	Number of Models
HGST (Western Digital)	10,384	5
Hitachi	12,991	9
Samsung	2	1
Seagate	36,213	24
Toshiba	247	3
Western Digital	3,061	36
Total	**62,898**	

The 2016 dataset from BackBlaze consists of 81,173 drives. The fraction of drives that failed in this period is 1.76 percent (1,431 drives). Table II shows the distribution in terms of the manufacturer, the number of drives, and the number of models for the 2016 dataset.

TABLE II: BackBlaze 2016 Dataset Distribution

Manufacturer	Number of Disks	Number of Models
HGST (Western Digital)	17,383	5
Hitachi	12,727	9
Samsung	1	1
Seagate	47,947	21
Toshiba	339	4
Western Digital	2,776	37
Total	**81,173**	

B. SMART Parameters Selection

To select the SMART parameters to build our model, we decided to use a methodology based on the following criteria:

a) *SMART parameter type (Raw and Normalized)*: Each attribute has a raw value, whose measurement is entirely up to the drive manufacturer (counts or a physical unit, such as degrees Celsius or seconds), and a normalized value, that transforms the raw value using a scale from 0 (bad) to some maximum (good) value. The maximum value is manufacturer and model dependent (different models from the same manufacturer can have a different initial value = 100, 200, or 253). We decided to use the RAW values of the selected parameters in our study because they were the most likely to remain the same among various makes and models.

b) *Statistical Analysis*: To keep the number of inputs to a manageable size and to eliminate irrelevant variables (parameters without any relationship to the target), we performed a statistical analysis of the 44 RAW parameters from the dataset using SAS Enterprise Miner. We applied the R-Square and Chi-Square methods to rank these 44 parameters based on their fitness for predicting disk failures while rejecting those with more than 90 percent of missing values. After this procedure, we were able to reduce the number of SMART parameters in our dataset to 21 variables by considering only the relevant parameters present in at least 10 percent of the disks in the studied dataset.

c) *Trend test*: We decided to perform a reverse arrangement test on the 21 SMART attributes selected from BackBlaze

dataset after applying the statistical analysis. We calculated the percentage of good and failed drives that have a trend per SMART parameter.

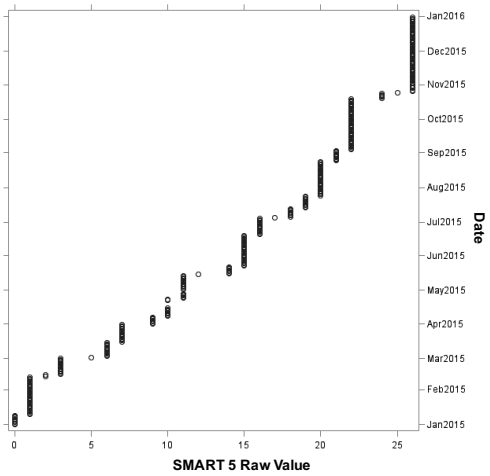

Fig. 1: Disk with a Trending Parameter

Figure 1 shows an example of a disk with a positive trending parameter (SMART 5 RAW), where the value of the studied variable increases as the time progresses.

Table III shows the results of the trend test. Based on these results, we selected the following parameters using the percentage of disks with a trend (high percentage of failed disk with a trend and low percentage of good disks with a trend):

- Reallocated Sectors Count (SMART 5): is the number of the bad sectors that have been found and remapped.
- Reported Uncorrectable Errors (SMART 187): is the number of errors that could not be recovered using hardware ECC.
- Command Timeout (SMART 188): is the number of aborted operations due to HDD timeout.
- Reallocation Event Count (SMART 196): is the number of attempts to transfer data from reallocated sectors to a spare area.
- Current Pending Sector Count (SMART 197): is the number of sectors waiting to be remapped because of unrecoverable read errors.
- Uncorrectable Sector Count (SMART 198): is the number of uncorrectable errors when reading/writing a sector.

Finally, it is important to mention that our six attributes include the five attributes used by BackBlaze to predict the failure or potential failure of a disk, showing that our trend test analysis agrees with the practical knowledge applied by BackBlaze to solve the hard disk failure detection problem.

C. Dataset Sampling

As we have seen, the 2015 BackBlaze dataset contained an overwhelming majority of disks that did not fail and a tiny minority of bad disks. Such unbalances make it harder for any machine learning technique to produce a predictor that could not neglect the failed disks.

Using our knowledge of the domain, we decided to solve this problem by using an informed undersampling method. We used the one-sided selection (OSS) method [16], to select a representative subset of the majority class E and combines it with the set of all minority instances $S_{\min}$ to form a preliminary set N, N=$E \cup S_{\min}$.

TABLE III: Reverse Arrangement Test SMART Parameter Distribution

Parameter	Description	%Good	%Bad	% Disks w/o Missing Values
Smart_5_Raw	Reallocated Sectors Count	1.887243	28.9916	100
Smart_187_Raw	Reported Uncorrectable Errors	1.562193	38.2227	56.69
Smart_188_Raw	Command Timeout	1.247511	10.27837	56.97
Smart_196_Raw	Reallocation Event Count	1.584494	27.32794	43.30
Smart_197_Raw	Current Pending Sector Count	1.417423	38.02521	100
Smart_198_Raw	Uncorrectable Sector Count	0.977863	27.10084	100

To implement the selected unbalanced dataset solution, we define $S_{\min}$ (minority class) as all instances of bad disks with at least one of the selected parameters with a trend, and E (majority class) as all instances of good disks with at least one of the selected parameters with a trend. Based on these definitions, we present the distribution of the instances for $S_{\min}$ and E:

Minority Class (Bad Disks):

- Total = 1,428.
- Bad disks with at least one parameter different than zero or missing = 891 (62.40 percent).
- Bad disks with all selected parameters equal to zero or missing = 537 (37.60 percent).

Majority Class (Good Disks):

- Total = 61,470.
- Good disks with at least one parameter different than zero or missing = 3,133 (5.09 percent).
- Good disks with all selected parameters equal to zero or missing = 58,337 (94.91 percent).

Our goal is to build the decision boundary between the minority class (failed disks) and the majority class (good disks), by only using 6.37 percent of the instances. The sole drawback of the approach is that it will classify as good all drives that have all six parameters equal to zero or missing.

D. Machine Learning Technique Selection

We implemented three different machine learning models (Decision Trees, Neural Networks, and Logistic Regression) using SAS Enterprise Miner version 14.1.

For the decision tree model, we used the following parameters:

- Max number of branches = 2.
- Max depth = 10.
- Assessment measure = Misclassification rate.

- (SAS automatically trains and prunes the tree based on the training and validation dataset).

For the neural network model, we used the following parameters:

- Architecture = Block Layers.
- Max number of interactions = 50.
- Max number of hidden units = 100.
- Activation functions = Direct, Logistic, Sine, Softmax, and Tanh.
- (SAS automatically selects the best network based on the selected parameters).

For the regression model, we used the following parameters:

- Main Effects.
- Two-factor Interactions.
- Polynomial terms (polynomial degree = 3).
- Regression type = Logistic.
- (SAS automatically selects the best regression model based on these parameters).

We decided to modify the prior probabilities of our training dataset to improve the capability of the proposed model to detect disk failures. The original training proportions were 0.22142 for the failed disks and 0.77858 for the good disks, and the new decision prior probabilities used are 0.5 for the minority class and 0.5 for the majority class. To validate our models, we used a five repetition ten-fold cross validation to provide ten random partitions of the original sample per repetition.

Model Comparison:

Tables IV and V present the results for the Fit Statistics and Confusion Matrix for the selected machine learning methods:

TABLE IV: Fit Statistics for the Selected Machine Learning Methods

Model	Misclassification Rate	Average Squared Error	ROC Index	Gini Coefficient
Decision Tree	0.15805	0.10036	0.883	0.766
Regression	0.22142	0.22142	0.5	0
Neural Network	0.22145	0.22145	0.5	0

TABLE V: Confusion Matrix for the Selected Machine Learning Methods

Model	False Negative	True Negative	False Positive	True Positive
Decision Tree	110	2,607	526	781
Regression	891	3,133	0	0
Neural Network	891	3,133	0	0

Out of these three methods, we selected the decision tree because it has the lowest misclassification rate (15.80 percent) and the highest number of true positives (781 failed disk predicted). The poor performance of both neural network and regression is a consequence of the presence of too many missing values for all instances of the minority class (disk failures). We then tried to impute the missing values using different parameters (median, mean, min, and max). Even though we saw an improvement for both neural network and regression after replacing the missing values with the median, the model with the best performance still was the decision tree.

Figure 2 shows the decision tree generated by SAS Enterprise Miner based on the selected parameters. For each node, we have the node id, the percentage of good disks (labeled as 0), the percentage of failed disks (labeled as 1), and the number of drives.

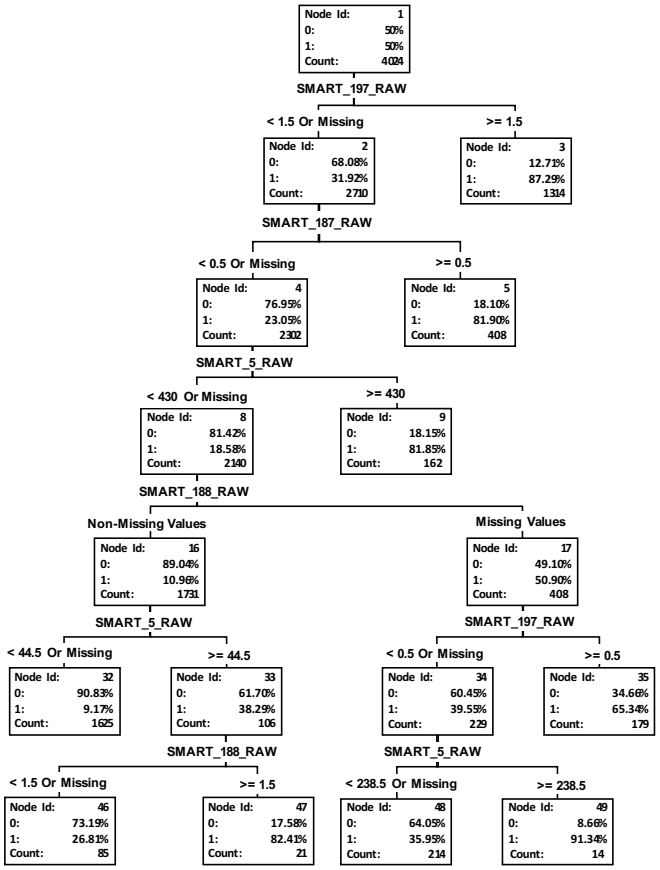

Fig. 2: Disk Failure Prediction Model based on Decision Trees

The limitation of the proposed model is shown by node 17. This node includes all instances with a zero or missing value for the SMART parameters 197, 187, 5 and 188, with a distribution of 49.10 percent of failed disks and 50.90 percent of good disks. This result shows that our model performs worst when an instance has all parameters equal to zero or missing.

IV. Experimental Results

We present the results of our study of the proposed model by using the 2015 and 2016 datasets from BackBlaze. We analyzed the failure detection rate (FDR), taking into consideration the fraction of predictable failures by using only the failed disks with at least one parameter greater than zero ($S_{\min}$ subset).

A. Testing with the complete 2015 Dataset

Since we built our model only 6.37 percent of the 2015 BackBlaze dataset, we had to check first how our model would perform on the whole 2015 BackBlaze dataset.

TABLE VI: Model Validation Results using the BackBlaze 2015 dataset

Failures Predicted	Failures Missed	False Alarms	Good Predicted
778	650	526	60,944

As we can see in table VI, our model achieves a failure detection rate (FDR) of 54.48 percent. To understand the behavior of the FDR for our model, we calculate the fraction of failures with non-zero values for one of our six indicators. From 1,428 failed disks in the dataset, only 891 have non-zero values, therefore the fraction of predictable failures is equal to 87.31 percent. This result shows that our model was able to predict almost 88 percent of the failures used to build the decision tree. For the false alarm rate (FAR), our model shows a value of 0.85 percent (because the number of true negatives is high).

B. Scoring with the 2016 BackBlaze Dataset

For the actual evaluation of the performance of our model, we used the 2016 dataset from BackBlaze.

TABLE VII: BackBlaze 2016 Scoring Results

Failures Predicted	Failures Missed	False Alarms	Good Predicted
747	684	551	79,191

As we can see in table VII, our model shows a failure detection rate (FDR) of 52.20 percent. For the studied dataset, only 858 failed disks have non-zero values. Therefore, the fraction of predictable failures is equal to 87.06 percent. This value can be explained by the relatively high number, 573 out of 1,431, that were not accompanied by any change in the values of any of our six SMART parameters. As our method was able to predict 747 failures out of the remaining 858 failures, it indeed predicted 87 percent of all failures that it could predict. For the false alarm rate (FAR), our model shows a value of 0.69 percent.

The failure detection rate (FDR) and the false alarm rate (FAR) are closely related to both predictable and unpredictable failures. Figure 3 shows the relationship between these parameters for the datasets.

To evaluate the timeliness of our predictions, we collected failure latencies, that is, the time interval between each failure prediction and the actual failure. A latency shorter than one or two days would give not enough time to save the information from a failed drive. Conversely, predicting disk failures too far ahead of time will offer little useful guidance to the system administrator.

Figure 4 shows the distribution of the predictions in terms of days. The results show that only 2.40 percent of the predictions occur within two days of the actual failure.

We also measured the latency in weeks to show how much ahead of time our model is making the predictions. Figure 5 shows the distribution of the predictions in terms of weeks.

These results show that 72.55 percent of disk failures are detected within four weeks of the actual failure date.

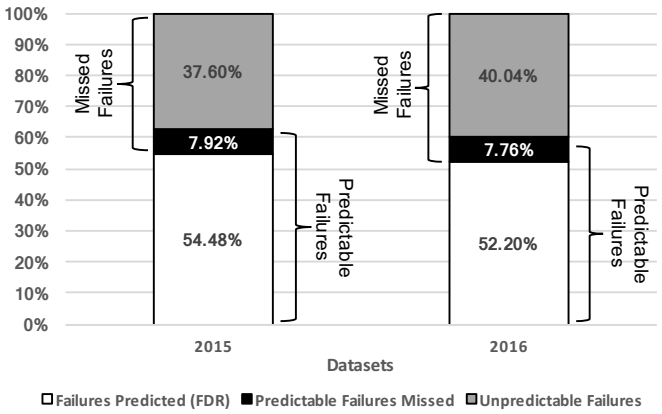

Fig. 3: Relationship between FDR and FAR

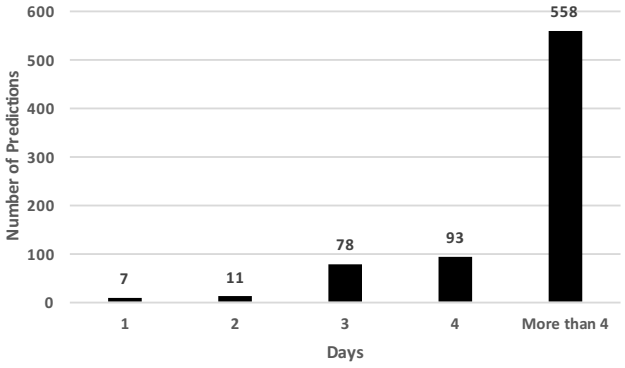

Fig. 4: Prediction Latency (Days)

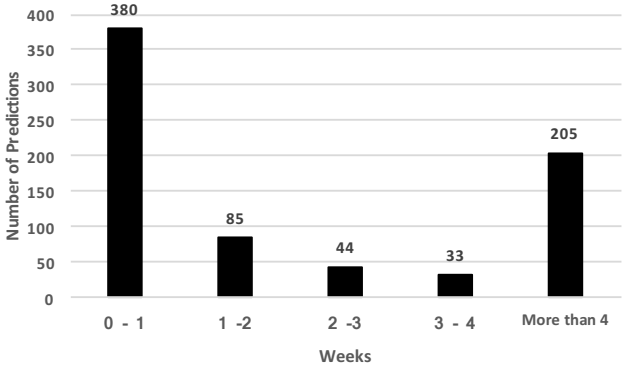

Fig. 5: Prediction Latency (Weeks)

C. Performance Comparison

We decided to compare the performance of our predictor against two extant predictors:

1) The disk failure predictor used in RAIDShield [9], which assumes that an excessive amount of sector errors typically precedes the failure of the whole-disk.
2) A baseline model used by BackBlaze that predicts a

disk failure each time any of the five relevant SMART parameters turns positive.

RAIDShield:

RAIDShield uses the SMART 5 Raw parameter (Reallocated Sector Count) to predict disks that are prone to failures. If the SMART 5 Raw exceeds a given failure threshold, the disk is considered to have become unreliable. The dataset they used to evaluate the performance of their detection mechanism is based on a population of 100,000 disks from the same family. Their results show a failure detection rate between 52 and 70 percent and a false alarm rate between 0.8 and 4.5 percent for SMART 5 values between 20 and 200.

We measure the performance of the RAIDShield predictor [9] using the 2016 dataset from BackBlaze. We used a range of values for the SMART 5 parameter between 1 and 600. We then compare these results against the results of our predictor for the same dataset.

Figure 6 shows the failure detection rate comparison between our predictor and RAIDShield detection mechanism. The results show that our predictor significantly outperforms RAIDShield detection mechanism with a difference between 36.82 percent (for SMART 5 value = 1) and 48.98 percent (for SMART 5 value = 600).

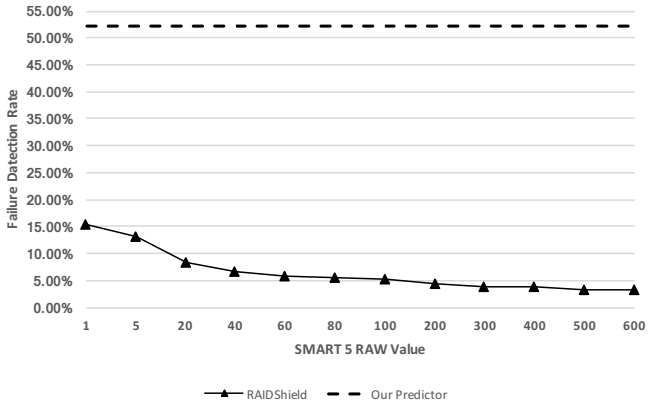

Fig. 6: Failure Detection Rate Comparison (Our Predictor vs RAIDShield)

Figure 7 shows the false alarm rate comparison between our predictor and RAIDShield detection mechanism. The very low FAR achieved by RAIDShield can be simply explained by observing that it makes much fewer predictions, good or bad, than our predictor.

These results are much worse than those reported by the authors of RAIDShield. This should not surprise us because RAIDShield was originally developed for and tested on a homogeneous population of disks, all the same make and model. In our study, only 18 percent of the failing disks were found to display any changes in the value of the SMART attribute associated with the number of relocated sectors before failing.

We conclude that our solution improves significantly the prediction of disk failures in heterogeneous environments when compared to RAIDShield while offering a similar performance for the true to false alarm ratio.

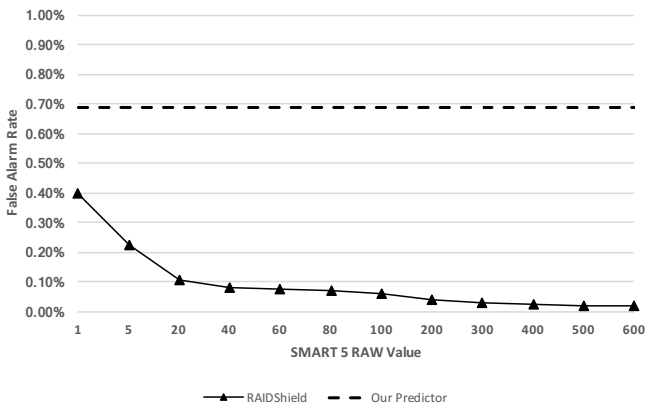

Fig. 7: False Alarm Rate Comparison (Our Predictor vs RAIDShield)

BackBlaze Predictor:

The original BackBlaze predictor uses five SMART parameters (SMART 5, SMART 187, SMART 188, SMART 197, and SMART 198). In other words, they use five out of our six parameters, missing SMART 196. They verify the status of a disk when the RAW value for one of these five attributes is greater than zero.

Since the baseline model predicts a disk failure each time any of our failure indicators turns positive, we can expect it to have a higher failure detection rate than our predictor. This was indeed the case: when we applied to the 2016 data, the baseline model correctly predicted 858 of the 1,431 observed disk failures, thus achieving a 60 percent failure detection rate, which is better than our predictor. This better performance came however with a price: the model issued 3,916 false alarms, that is, slightly more than 4.5 false alarm per correct prediction. Table VIII shows the results of the scoring process for the 2016 BackBlaze dataset with the baseline model:

TABLE VIII: Scoring the Baseline model with BackBlaze 2016 dataset

Failures Predicted	Failures Missed	False Alarms	Good Predicted
858	573	3,916	75,826

In addition, we observed that 25 percent of its predictions were made more than sixty days ahead of time of the actual disk failure. This very long delay would greatly complicate the task of asserting the correctness of the predictions made by the model. It could indeed result in having many of these long-term predictions incorrectly classified as false alarms. Should this be the case, the effective failure prediction rate of the model would fall below the failure detection rate of our predictor.

V. Conclusion

We have presented a decision-tree based disk failure predictor for heterogeneous populations of disks, such as the ones

encountered in many large data centers. The main problem we encountered was the lack of standardization among the SMART attributes of various makes and models of the disk population we investigated. As a result, we were only able to identify six SMART attributes that were the strongly correlated with an impending disk failure for our whole disk population. We found out that only a small minority of disks (around 3 percent) were flagged by one or more of these attributes. We then used machine learning to build a decision tree that would separate true failure predictions from false alarms among the flagged disks. Our decision tree was built using disk reliability data collected at BackBlaze in 2015 and scored using data collected there during 2016.

Our results indicate that we can predict 52 percent of all disk failures. This corresponded to 87 percent of the disk failures that were preceded by any change in one of this SMART attributes we monitored.

We also compared the performance of our predictor with that of the RAIDShield predictor and that of a baseline prediction model that predicted a failure each time one of our failure indicators turned positive. We found out that RAIDShield performed very poorly in our heterogeneous environment and was never able to predict more than 18 percent of disk failures. While the BackBlaze predictor (baseline model) could predict 60 percent of disk failures, it also generated between 4 and 5 false alarms per correct prediction. These results illustrate the difficulty of developing a good generic disk failure predictor for a heterogeneous population of disks.

In a future work, we plan to regroup disks by manufacturer and construct a separate disk failure predictor for each make. This should allow us to include more SMART attributes in our decision trees and, hopefully, result in higher failure detection rates. Even then, our generic predictor would still apply to makes of disks that are barely present in the disk population or were recently inserted.

REFERENCES

[1] E. Pinheiro, W.-D. Weber, and L. A. Barroso, "Failure trends in a large disk drive population," in *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, FAST '07, (Berkeley, CA, USA), pp. 17–29, USENIX Association, 2007.

[2] B. Schroeder and G. A. Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?," in *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, FAST '07, (Berkeley, CA, USA), pp. 1–16, USENIX Association, 2007.

[3] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in *Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data*, SIGMOD '88, (New York, NY, USA), pp. 109–116, ACM, 1988.

[4] T. J. E. Schwarz and W. A. Burkhard, "RAID organization and performance," in *Proceedings of the 12th International Conference on Distributed Computing Systems*, pp. 318–325, Jun 1992.

[5] C. Xu, G. Wang, X. Liu, D. Guo, and T. Y. Liu, "Health status assessment and failure prediction for hard drives with recurrent neural networks," *IEEE Transactions on Computers*, vol. 65, pp. 3502–3508, Nov 2016.

[6] M. M. Botezatu, I. Giurgiu, J. Bogojeska, and D. Wiesmann, "Predicting disk replacement towards reliable data centers," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 39–48, ACM, 2016.

[7] J. Li, X. Ji, Y. Jia, B. Zhu, G. Wang, Z. Li, and X. Liu, "Hard drive failure prediction using classification and regression trees," in *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pp. 383–394, June 2014.

[8] "BackBlaze quarterly hard drive reliability reports." https://www.backblaze.com/b2/hard-drive-test-data.html. Accessed: 2016-11-20.

[9] A. Ma, F. Douglis, G. Lu, D. Sawyer, S. Chandra, and W. Hsu, "RAIDShield: Characterizing, monitoring, and proactively protecting against disk failures," in *13th USENIX Conference on File and Storage Technologies (FAST 15)*, (Santa Clara, CA), pp. 241–256, USENIX Association, 2015.

[10] B. Zhu, G. Wang, X. Liu, D. Hu, S. Lin, and J. Ma, "Proactive drive failure prediction for large scale storage systems," in *2013 IEEE 29th Symposium on Mass Storage Systems and Technologies (MSST)*, pp. 1–5, May 2013.

[11] F. Salfner, M. Lenk, and M. Malek, "A survey of online failure prediction methods," *ACM Comput. Surv.*, vol. 42, pp. 10:1–10:42, Mar. 2010.

[12] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Machine learning methods for predicting failures in hard drives: A multiple-instance application," *J. Mach. Learn. Res.*, vol. 6, pp. 783–816, Dec. 2005.

[13] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Hard drive failure prediction using non-parametric statistical methods," in *Proceedings of ICANN/ICONIP*, 2003.

[14] G. Hamerly and C. Elkan, "Bayesian approaches to failure prediction for disk drives," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, (San Francisco, CA, USA), pp. 202–209, Morgan Kaufmann Publishers Inc., 2001.

[15] J. Li, R. J. Stones, G. Wang, Z. Li, X. Liu, and K. Xiao, "Being accurate is not enough: New metrics for disk failure prediction," in *2016 IEEE 35th Symposium on Reliable Distributed Systems (SRDS)*, pp. 71–80, Sept 2016.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng.*, vol. 21, pp. 1263–1284, Sept. 2009.